概要

相対的意味論に基づく変換主導型統計機械翻訳 TDSMT[1](Transfer Driven Statistical Machine Translation : 以下 TDSMT と記述する) が提案されている. 本研究では TDSMT を従来手法として扱う. 従来手法は変換テーブルを用いて翻訳を行う. 変換テーブルは学習文対から自動作成する. 学習文対は英語文と日本語文の対である. また, 変換テーブルは “「A が B」ならば「C は D」” で表現する. しかし, 従来手法は自動で変換テーブルを作成するため, 誤った変換テーブルを作成する場合がある. そこで, 本研究は変換テーブルの A, B, C, D の文中における前後環境を利用して誤った変換テーブルを削除する方法を提案する.

本研究は誤った変換テーブルの削除を目的とする. 変換テーブルは A と C, また B と D の置き換えを想定している. 正しい変換テーブルは A と C, また B と D の文中での置き換えが可能である場合が多い. そこで本手法は, 文中における句の置き換え可否を検証し, 誤った変換テーブルを削除する. 具体的には, 前後の単語の比較を学習文対内にて行い文中における句の置き換え可否を検証する.

実験では, 従来手法によって作成した変換テーブルに対して提案手法を行った. そして, 提案手法を行なう前と後の変換テーブルの数と精度を比較した. 実験の結果, 変換テーブルの数は提案手法を行った後, 減少した. しかし, 提案手法によって変換テーブルの精度が向上することが確認できた.

目 次

図 目 次

表 目 次

第1章　はじめに

機械翻訳には様々な手法が存在する. 現在は統計機械翻訳やニューラル機械翻訳が主流である. しかし, 統計機械翻訳は高精度の翻訳結果を出力するために多量のパラレルコーパスを必要とする. ヨーロッパ圏の言語は多量のパラレルコーパスが Europarl Corpus[2] によって存在するが, 日英パラレルコーパスは Europarl Corpus と比較して数が不足している. また, ニューラル機械翻訳は出力の導出過程の解析が困難である.

機械翻訳において, 相対的意味論に基づく変換主導型統計機械翻訳 TDSMT が提案されている. 従来手法は変換テーブルを用いて, 学習文対を変換し, 出力文を作成する. 変換テーブルは “「A が B」ならば「C は D」” で表現する. 変換テーブルは学習文対 (パラレルコーパス) から自動作成する. 作成には IBM Model 1[3], 単語レベル文パターンを利用する. 従来手法は学習文対 1 対から複数の変換テーブルを作成する. また, 出力の導出過程の解析もニューラル機械翻訳と比べ, 容易である.

従来手法は学習文対の単語を変換テーブルを利用し, 置き換えることによって出力文を得る. つまり, 変換テーブルは A と C, そして B と D が置き換え可能な関係が想定される. しかし, 従来手法は自動で変換テーブルを作成するため, 誤った変換テーブルを作成する場合がある. 誤った変換テーブルとは, 文中において A と C, そして B と D の置き換えが不可能である変換テーブルを意味する. そこで, 本研究は誤った変換テーブルの削除を目的とする. 本研究では提案手法として, 前後環境を利用して誤った変換テーブルを削除する方法を提案する.

第2章　従来研究

2.1　統計翻訳

本節は西尾ら [4] の抜粋である.

2.1.1　概要

統計翻訳とは, 機械翻訳手法の一種である. 原言語と目的言語の対訳文を大量に収集した対訳文より, 自動的に翻訳規則を獲得し翻訳を行う.
統計翻訳には単語に基づく統計翻訳と句に基づく統計翻訳があり, 初期の統計翻訳では単語に基づく統計翻訳が用いられていたが, 翻訳精度は高くなかった. しかし近年, 句に基づく統計翻訳が提案され, 単語に基づく統計翻訳に比べて翻訳精度が高いことがわかった. このため現在は句に基づく統計翻訳が主流となっている.

2.1.2　単語に基づく統計翻訳

単語に基づく統計翻訳は単語対応の翻訳モデルを用いている. 例として, ある日本語文を英語文に翻訳する場合を考える. 日本語単語を英語に翻訳し, 日本語単語の語順と同じ並びで英単語を並べて翻訳する. 単語に基づく統計翻訳は単語対応の確率を得る IBM 翻訳モデルが用いられている.

2.1.3　IBM 翻訳モデル

IBM 翻訳モデルを以下に示す. 本節は力久ら [5] の抜粋である. 統計翻訳の代表的なモデルとして, IBM の Brown らによる仏英翻訳モデルがある.IMB 翻訳モデルは, 単語に基づく統計翻訳を想定して作成された, 単語対応の確率モデルである. この翻訳モデルは順に複雑な計算を行うモデル 1 から 5 の 5 つのモデルで構成される.

本章では. 原言語であるフランス語文を F, 目的言語である英語文を E として定義する.

IBM モデルでは, フランス語文 E, 英語文 F の翻訳モデル $P(F|E)$ を計算するために, アライメント a を用いる. 以下に IBM モデルの基本式を示す.

$$P(F|E) = \sum_{a} P(F, a|E) \tag{2.1}$$

アライメントとは仏単語と英単語の対応を意味している.IBM モデルのアライメントでは, 各仏単語 f に対応する英単語 e は 1 つあり, 各英単語 e に対応する仏単語は 0 から n 個ある. また仏単語 f において適切な英単語と対応しない場合, 英語文の先頭に空単語 e_0 があると仮定し, その仏単語 f と空単語 e_0 を対応づける.

・モデル 1

(2.1) 式は以下の式に分解することができる.m はフランス語文の長さ, a_1^{j-1} はフランス語文における, 1 番目から $j-1$ 番目までのアライメント, f_1^{j-1} はフランス語文における, 1 番目から $j-1$ 番目まで単語を表している.

$$P(F, a|E) = P(m|E) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) P(f_j|a_1^{j}, f_1^{j-1}, m, E) \tag{2.2}$$

(2.2) 式ではとても複雑であるので計算が困難である. そこで, モデル 1 では以下の仮定により, パラメータの簡略化を行う.

- フランス語文の長さの確率 ϵ は m, E に依存しない

$$P(m|E) = \epsilon$$

- アライメントの確率は英語文の長さ l に依存する

$$P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) = (l+1)^{-1}$$

- フランス語の翻訳確率 $t(f_j|e_{a_j})$ は, 仏単語 f_j に対応する英単語 e_{a_j} に依存する

$$P(f_j|a_1^{j}, f_1^{j-1}, m, e) = t(f_j|e_{a_j})$$

パラメータの簡略化を行うことで, $P(F, a|E)$ と $P(F, E)$ は以下の式で表される.

$$P(F, a|E) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.3}$$

$$P(F|E) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.4}$$

$$= \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) \tag{2.5}$$

モデル 1 では翻訳確率 $t(f|e)$ の初期値が 0 以外の場合, Expectation-Maximization(EM) アルゴリズムを繰り返し行うことで得られる期待値を用いて最適解を推定する.EM アルゴリズムの手順を以下に示す.

手順 1 翻訳確率 $t(f|e)$ の初期値を設定する.

手順 2 仏英対訳対 $(F^{(s)}, E^{(s)})$(但し, $1 \leq s \leq S$) において, 仏単語 f と英単語 e が対応する回数の期待値を以下の式により計算する.

$$c(f|e; F, E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{2.6}$$

$\delta(f, f_j)$ はフランス語文 F 中で仏単語 f が出現する回数, $\delta(e, e_i)$ は英語文 E 中で英単語 e が出現する回数を表している.

手順 3 英語文 $E^{(s)}$ の中で 1 回以上出現する英単語 e に対して, 翻訳確率 $t(f|e)$ を計算する.

1. 定数 λ_e を以下の式により計算する.

$$\lambda_e = \sum_{f} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) \tag{2.7}$$

2. (2.7) 式より求めた λ_e を用いて, 翻訳確率 $t(f|e)$ を再計算する.

$$t(f|e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)}) = \frac{\sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})}{\sum_{f} \sum_{s=1}^{S} c(f|e; F^{(s)}, E^{(s)})} \tag{2.8}$$

手順 4 翻訳確率 $t(f|e)$ が収束するまで手順 2 と手順 3 を繰り返す.

・モデル 2

モデル 1 では, 全ての単語の対応に対して, 英語文の長さ l にのみ依存し, 単語対応の確率を一定としている. そこで, モデル 2 では, j 番目の仏単語 f_j と対応する英単語の位置 a_j は英語文の長さ l に加えて, j と, フランス語文の長さ m に依存し, 以下のような関係とする.

$$a(a_j|j,m,l) \equiv P(a_j|a_1^{j-1}, f_1^{j-1}, m, l) \tag{2.9}$$

この関係からモデル 1 における (2.4) 式は, 以下の式に変換できる.

$$\begin{aligned} P(F|E) &= \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) a(a_j|j,m,l) && (2.10) \\ &= \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_{a_j}) a(a_j|j,m,l) && (2.11) \end{aligned}$$

モデル 2 では, 期待値は $c(f|e;F,e)$ と $c(i|j,m,l;F,E)$ の 2 つが存在する. 以下の式から求められる.

$$\begin{aligned} c(f|e;F,E) &= \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \sum_{j=1}^{m} \delta(f,f_j) \sum_{i=1}^{l} \delta(e,e_i) && (2.12) \\ &= \sum_{j=1}^{m} \sum_{i=0}^{l} \frac{t(f|e) a(i|j,m,l) \delta(f,f_j) \delta(e,e_i)}{t(f|e_0) a(0|j,m,l) + \cdots + t(f|e_l) a(l|j,m,l)} && (2.13) \\ c(i|j,m,l;F,E) &= \sum_{a} P(a|E,F) \delta(i,a_j) && (2.14) \\ &= \frac{t(f_j|e_i) a(i|j,m,l)}{t(f_j|e_0) a(0|j,m,l) + \cdots + t(f_j|e_l) a(l|j,m,l)} && (2.15) \end{aligned}$$

$c(f|e;F,E)$ は対訳文中の英単語 e と仏単語 f が対応付けされる回数の期待値, $c(i|j,m,l;F,E)$ は英単語の位置 i が仏単語の位置 j に対応付けされる回数の期待値を表している.

モデル 2 では, EM アルゴリズムで計算すると複数の極大値が算出され, 最適解が得られない可能性がある. モデル 1 では $a(i|j,m,l) = (l+1)^{-1}$ となるモデル 2 の特殊な場合であると考えられる. したがって, モデル 1 を用いることで最適解を得ることができる.

・モデル 3

モデル 3 は, モデル 1 とモデル 2 とは異なり, 1 つの単語が複数対応する単語の繁殖数や単語の翻訳位置の歪みについて考慮する. またモデル 3 では単語の位置を絶対位置とし

て考える. モデル 3 では以下のパラメータを用いる.

- 翻訳確率 $P(f|e)$
  英単語 e が仏単語 f に翻訳される確率
- 繁殖確率 $n(\phi|e)$
  英単語 e が ϕ 個の仏単語と対応する確率
- 歪み確率 $d(j|i,m,l)$
  英語文の長さ l, フランス語文の長さ m のとき, i 番目の英単語 e_i が j 番目の仏単語 f_j に翻訳される確率

さらに, 英単語が仏単語に翻訳されない個数を ϕ_0 とし, その確率 p_0 を以下の式で求める. このとき, 歪み確率は $\frac{1}{\phi_0!}$ で, $p_0 + p_1 = 1$ で p_0, p_1 は 0 より大きいとする.

$$P(\phi_0|\phi_1^l, E) = \left(\begin{array}{c} \phi_1 + \cdots + \phi_l \\ \phi_0 \end{array} \right) p_0^{\phi_1+\cdots+\phi_l-\phi_0} p_1^{\phi_0} \tag{2.16}$$

したがって, モデル 3 は以下の式で求められる.

$$\begin{aligned} P(F|E) &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} P(F, a|E) && (2.17) \\ &= \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \left(\begin{array}{c} m - \phi_0 \\ \phi_0 \end{array} \right) p_0^{m-2\phi_0} p_1^{\phi_0} \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \\ &\times \prod_{j=1}^{m} t(f_j|e_{a_j}) d(j|a_j, m, l) && (2.18) \end{aligned}$$

モデル 3 では, 全てのアライメントを計算するため, 計算量が膨大となるので期待値を近似により求める.

・モデル 4

モデル 4 では, モデル 3 と異なり, 単語の位置を絶対位置ではなく, 相対位置で考える. またモデル 3 では考慮されていない各単語の位置, 例えば形容詞と名詞の関係を考慮する. モデル 4 では歪み確率 $d(j|i.m,l)$ を 2 つの場合で考える.

- 繁殖数が 1 以上である英単語に対応する仏単語の中で, 最も文頭に近い場合

$$P(\Pi_{[i]1} = j|\pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_1(j - \odot_{i-1}|\mathcal{A}(e_{[i-1]}), \mathcal{B}(f_j)) \tag{2.19}$$

$\odot_{i-1}$ は i-1 番目の英単語が対応する仏単語の位置を表している.

- それ以外の場合

$$P(\Pi_{[i]k} = j|\pi_{[i]1}{}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) = d_{>1}(j - \pi_{[i]k-1}|\mathcal{B}(f_j)) \tag{2.20}$$

$\pi_{[i]k-1}$ は同じ英単語に対応している直前の仏単語を表している.

・モデル 5

モデル 4 では, 単語の位置に関して直前の単語以外は考慮されていない. したがって, 複数の単語が同じ位置に生じたり, 単語の存在しない位置が生成される. モデル 5 では, この問題を避けるために, 単語を空白部分に配置するよう改善が施されている.

- 繁殖数が 1 以上である英単語に対応する仏単語の中で, 最も文頭に近い場合

$$\begin{aligned} P(\Pi_{[i]1} &= j|\pi_1^{[i]-1}, \tau_0^l, \phi_o^l, E) \\ &= d_1(v_j|\mathcal{B}(f_j), v_{\odot_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1})) \end{aligned}$$

v_j は j 番目までの空白数, $\mathcal{A}$ は英語の単語クラス $\mathcal{B}$ はフランス語の単語クラスを表している.

- それ以外の場合

$$\begin{aligned} P(\Pi_{[i]k} &= j|\pi_{[i]1}{}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, E) \\ &= d_{>1}(v_j - v_{\pi_{[i]k-1}}|\mathcal{B}(f_j), v_m - v_{\pi_{[i]k-1}} - \phi_{[i]} + k)(1 - \delta(v_j, v_{j-1})) \end{aligned}$$

2.1.4 GIZA++

GIZA++[8] とは, 統計翻訳で用いることを前提に作られたツールである.IBM 翻訳モデルを用いて, 対訳文 (原言語文と目的言語文の対) から対訳単語と単語翻訳確率を自動的に得る.

2.2 句に基づく統計翻訳

句に基づく統計翻訳は句対応の翻訳モデルを用いる. 原言語文を目的言語文に翻訳する場合に, 隣接する複数の単語 (フレーズ) を用いて翻訳を行う方法である. 本研究では日

英方向の翻訳を行うため, 日英統計翻訳を説明する. 日英統計翻訳システムの枠組みを図 2.1 に示す.

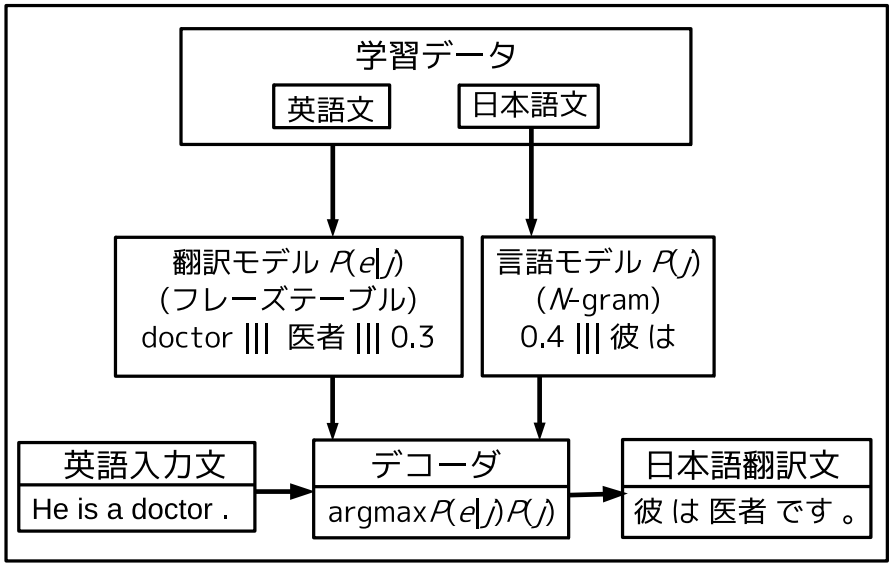

図 2.1: 日英統計翻訳の枠組み

$$E = argmax_j P(e|j) \tag{2.21}$$

$$\simeq argmax_j P(j|e)P(e) \tag{2.22}$$

ここで $P(j|e)$ は翻訳モデル, $P(e)$ は言語モデルを示す.$P(e)$ が単語であれば “単語に基づく統計翻訳” のモデル, $P(e)$ が句であれば, “句に基づく統計翻訳” のモデルとなる.

また, 学習データとは対訳文 (英語文と日本語文の対) を大量に用意したものである. 学習データに含まれる各々のデータから, 翻訳モデルと言語モデルを学習する.

2.2.1 翻訳モデル

翻訳モデルとは, 膨大な量の対訳データを用いて英語のフレーズが日本語のフレーズへ確率的に翻訳を行うためのモデルである. この翻訳モデルはフレーズテーブルで管理されている. 以下にフレーズテーブルの例を示す.

フレーズテーブルの例

The flower ||| その 花 ||| 0.428571 0.0889909 0.428571 0.0907911 2.718
Tonight’s concert is ||| 今晩 の コンサート は ||| 0.5 0.000223681 0.5 0.0124601 2.718

左から英語フレーズ, 日本語フレーズ, フレーズの英日方向の翻訳確率 $P(j|e)$, 英日方向の単語の翻訳確率の積, フレーズの日英方向の翻訳確率 $P(e|j)$, 日英方向の単語の翻訳確率の積, フレーズペナルティ (値は常に自然対数の底 e=2.718) である.

2.2.2 フレーズテーブル作成法

まず, GIZA++を用いて学習文から英日, 日英方向の双方向で最尤な単語アライメントを得る. 英日方向の単語対応の例を表 2.1, 日英方向の単語対応の例を表 2.2 に示す. また, ●は単語が対応した箇所を示す.

表 2.1: 日英方向の単語対応

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

表 2.2: 英日方向の単語対応

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に						
行っ		●				
た						

次に, 得られた双方向の単語アライメントを用いて, 複数単語のアライメントを得る. このアライメントは双方向の単語対応の和集合と積集合から求める. ヒューリスティックスとして双方向ともに対応する単語対応を用いる “intersection”, 双方向のどちらか一方でも対応する単語対応を全て用いる “union” がある. 表 2.1 と表 2.2 を用いた “intersection” の例を表 2.3, に “union” の例を表 2.4 に示す.

表 2.3: intersection の例

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に						
行っ		●				
た						

表 2.4: union の例

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

また “intersection” と “union” の中間のヒューリスティックスとして “grow” と “grow-diag” がある. これら 2 つのヒューリスティックスでは “intersection” の単語対応と “union” の単語対応を用いる.“grow” は縦横方向, “grow-diag” は縦横対角方向に, “intersection” の単語対応から “union” の単語対応が存在する場合にその単語対応も用いる.“grow-diag” の例を表 2.5 に示す.

表 2.5: grow-diag の例

	He	went	to	kyoto	on	business
彼	●					
は						
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

“grow-diag” の最後に行う処理として “final” と “final-and” がある.“final” は少なくとも片方の言語の単語対応がない場合に, “union” の単語対応を追加する. また, “final-and” は, 両側言語の単語対応がない場合に, “union” の候補対応点を追加する.“grow-diag-final-and” の例を表 2.6 に示す.

表 2.6: grow-diag-final-and の例

	He	went	to	kyoto	on	business
彼	●					
は			●			
仕事						●
で					●	
京都				●		
に		●			●	
行っ		●				
た		●				

得られた単語アライメントから, 全ての矛盾しないフレーズ対を得る. このとき, そのフレーズ対に対して翻訳確率を計算し, フレーズ対に確率値を付与することでフレーズテーブルを作成する.

2.2.3 言語モデル

言語モデルとは, 人間が用いる言葉の自然な並びを確率としてモデル化したものであり, 膨大な量の単言語データを用いて単語の列や文字の列が起こる遷移確率を付与したものである. 言語モデルには以下のようなものがある.

N-gram(2.23)

統計翻訳では主に N-gram を用いる. tri-gram の式を式 2.23 に示す.

$$\sum_{i=0}^{N-1} \log_2 \frac{count(E_{i-2}, E_{i-1}, E_i)}{count(E_{i-2}, E_{i-1})} \tag{2.23}$$

E_i : 英語単語　　N : 英文の単語数
C : 対訳学習文の頻度

実際の計算例を (2.24) に示す.

$$\begin{aligned} &\log_2 P(I\ have\ a\ dog.) \\ &= \log_2 \frac{count(I\ have\ a)}{count(I\ have)} \\ &+ \log_2 \frac{count(have\ a\ dog)}{count(have\ a)} \\ &+ \log_2 \frac{count(a\ dog.)}{count(a\ dog)} \\ &= \log_2 \frac{140}{1,007} + \log_2 \frac{2}{465} + \log_2 \frac{14}{31} \\ &= -11.8545 \end{aligned} \tag{2.24}$$

High order Joint Probability(2.25)

本研究では, 言語モデルに Tri-gram の代わりに High order Joint Probability を使用する. High order Joint Probability を式 2.25 に示す.

$$\sum_{j=0}^{M-1}\sum_{i=0}^{N-1} count(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i) \times \log_2 \frac{count(J_{j-2}, J_{j-1}, J_j, E_{i-2}, E_{i-1}, E_i)}{count(J_{j-2}, J_{j-1}, J_j)count(E_{i-2}, E_{i-1}, E_i)} \tag{2.25}$$

J_j : 日本語単語　M : 日本語文の単語数
E_i : 英語単語　　N : 英文の単語数
P : 出現確率

実際の計算例を (2.26) に示す. また, 計算式が長くに及ぶため, 第 1 項のみ計算例を示す.

$$\begin{aligned} &P(\text{ぶらんこが揺れている。　}\mathit{The\ swing\ is\ swinging.}) \\ &= count(\text{ぶらんこ が　}\mathit{The\ swing}) log_2 \frac{count(\text{ぶらんこ が　}\mathit{The\ swing})}{count(\text{ぶらんこ が})P(\mathit{The\ swing})} + ... \\ &= \frac{1}{100,000} log_2 \frac{\frac{1}{100,000}}{\frac{2}{100,000}\frac{1}{100,000}} + ... \end{aligned} \tag{2.26}$$

High order Dice(2.27)

$$\sum_{j=0}^{M-1}\sum_{i=0}^{N-1}\log_2\frac{2\cdot count(J_{j-2},J_{j-1},J_j,E_{i-2},E_{i-1},E_i)}{count(J_{j-2},J_{j-1},J_j)+count(E_{i-2},E_{i-1},E_i)} \tag{2.27}$$

実際の計算例を (2.28) に示す. また, 計算式が長くに及ぶため, 第 1 項のみ計算例を示す.

$$\begin{aligned} &P(\text{ぶらんこが揺れている。}\quad The\ swing\ is\ swinging.)\\ &=\log_2\frac{2\cdot count(\text{ぶらんこ が}\quad The\ swing}{count(\text{ぶらんこ が})+count(The\ swing)}+...=\frac{2\cdot\frac{1}{100{,}000}}{\frac{2}{100{,}000}+\frac{1}{100{,}000}}+... \end{aligned} \tag{2.28}$$

High order Log Linear(2.29)

$$\sum_{j=0}^{M-1}\sum_{i=0}^{N-1}\log_2\left\{\frac{count(J_{j-2},J_{j-1},J_j,E_{i-2},E_{i-1},E_i)}{count(J_{j-2},J_{j-1},J_j)}\times\frac{count(E_{i-2},E_{i-1},E_i,J_{j-2},J_{j-1},J_j)}{count(E_{i-2},E_{i-1},E_i)}\right\} \tag{2.29}$$

実際の計算例を (2.30) に示す. また, 計算式が長くに及ぶため, 第 1 項のみ計算例を示す.

$$\begin{aligned} &P(\text{ぶらんこが揺れている。}\quad \textit{The swing is swinging.})\\ &=\log_2\left\{\frac{count(\text{ぶらんこ が}\quad \textit{The swing})}{count(\text{ぶらんこ が})}\times\frac{count(\textit{The swing}\quad \text{ぶらんこ が})}{count(\textit{The swing})}\right\}\\ &=\log_2\left\{\frac{\frac{1}{100{,}000}}{\frac{2}{100{,}000}}\times\frac{\frac{1}{100{,}000}}{\frac{1}{100{,}000}}\right\} \end{aligned} \tag{2.30}$$

2.2.4 デコーダ

デコーダは, 翻訳モデルと言語モデルを用いて, 確率が最大となる翻訳候補を探索し, 出力を行う変換器のことである. 代表的なデコーダとして, “Moses” [6] がある.

入力文として “She is a teacher .” が与えられたときの翻訳例を図 2.2 に示す.

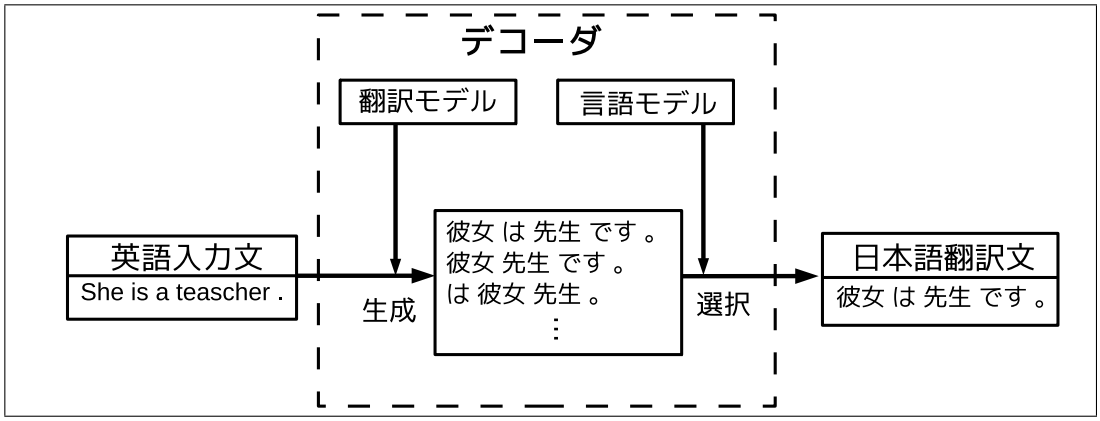

図 2.2: デコーダの動作例

日英統計翻訳において, $argmax_e P(e|j)P(j)$ の確率が最大となる英語文を出力するために, 適切な順序で日本語と英語の単語対応を得る必要がある. しかし, 適切な日本語文を決定するためには, 計算量が膨大となり, かつ莫大な時間が必要となる. そこで計算量を削減するために, ビームサーチ法を用いる.

ビームサーチ法とは, 翻訳候補の探索において, 翻訳確率の低い翻訳候補を枝刈りし, 探索範囲を減退する方法である. 探索領域の中で一定の確率以上の翻訳候補のみを残し, それ以外の翻訳候補は除外する.

ただし, ビームサーチ法は, 切り捨てられた翻訳候補が文章全体で見たときに, 最大の確率を持つ翻訳候補であったという可能性がある. そのため選択した翻訳文が最適解であるとは限らないという問題がある.

2.3 相対的意味論に基づく変換主導統計機械翻訳 (TDSMT)

本章は中村 [7] らの抜粋である. “相対的意味論に基づく変換主導型統計機械翻訳 (TDSMT)” は, 安場らが提案した機械翻訳の手法である. TDSMT は, 学習文対と, 変換テーブルを用いて, 原言語文を入力とし, 目的言語文を出力する. 変換テーブルは “*A* が *B* ならば *C* は *D*” で表現する. *A* は学習文対中の原言語句, *B* は学習文対中の目的言語句, *C* は入力文中の原言語句, *D* は出力文中の目的言語句である.

原言語入力文が, 学習文対の原言語側と一致するまで, 入力文と変換テーブル中の *AC* を照合する. 次に, 一致した学習文対の目的言語側を, 照合した変換テーブルの *BD* に従って変換し, 目的言語翻訳文を出力する.

TDSMT は変換テーブルを学習文対から自動作成する. しかし, 問題点として, 誤った対訳を含む変換テーブルを作成することがあげられる.

2.3.1 TDSMT の手順

TDSMT の手順を示す. 手順は “学習” と “翻訳” の二部からなる.

2.3.2 学習の手順

TDSMT における学習は “変換テーブルの作成” のみである. 本節で作成手順を示す.

手順 1 対訳単語の作成

学習文対と対訳単語確率 (IBM Model 1) を利用して, 対訳単語を作成する. このとき付与される対訳単語確率を P_w とする. 例として, 表 2.7 に示す学習文対を使用して, 表 2.8 に示す対訳単語を作成する. 表 2.8 の値は例であり, 実際の数値とは異なる.

表 2.7: 対訳単語作成に用いる学習文対

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.

表 2.8: 作成される対訳単語

	日本語単語	英語単語	p_w
対訳単語 1	彼	His	0.4
対訳単語 2	弟	brother	0.7
対訳単語 3	学生	student	0.6

手順 2　単語レベル文パターンの作成

学習文対内で対訳単語に当たる部分を変数化し, 単語レベル文パターンを作成する. 例を表 2.9 に示す.

表 2.9: 単語レベル文パターンの作成例

学習文対 (日本語側)	彼の兄は医者だ。
学習文対 (英語側)	His brother is a doctor.
単語レベル文パターン (日本語側)	*X0* の *X1* は *X2* だ
単語レベル文パターン (英語側)	*X0 X1* is a *X2*

手順 3　変換テーブルの作成

学習文対と単語レベル文パターンを照合する. 変数化した対訳単語と, 変数に当たる対訳句を変換テーブルとする. 表 2.10 では変数 *N2* の部分から変換テーブル “「学生」が「student」ならば「教師」は「teacher」” が得られる.

表 2.10: 変換テーブルの作成例

学習文対 (日本語側)	彼の弟は学生だ。
学習文対 (英語側)	His brother is a student.
単語レベル文パターン (日本語側)	*X0* の *X1* は *X2* だ。
単語レベル文パターン (英語側)	*X0 X1* is a *X2* .
照合する学習文対 (日本語側)	私の母は教師だ。
照合する学習文対 (英語側)	My mother is a teacher.
変換テーブル *(X2)*	*A*:学生　*B*:student *C*:教師　*D*:teacher

手順 4　変換テーブルに確率を付与

対訳単語確率 P_w を利用し, 変換テーブルに確率を付与する. この確率を変換テーブル確率 P_v とする.

1. 変換テーブルの *CD* に存在する全ての日英単語の組み合わせを確認する.
2. 日本語単語に対応する英語単語の中で, 対訳単語確率 P_w の最大値を得る.
3. 各日本語単語について得られた値と, 変換テーブルの *AB* の対訳単語確率 P_w について, 対数の総和を求める.

2.3.3 翻訳の手順

本節で TDSMT における翻訳の手順を示す. 入力文を「私の姉は教師だ。」とする.

手順 1　入力文に日本語側の変換テーブルを適用

変換テーブルの C と A を利用して, 入力文を学習文対の日本語側と一致させる. 表 2.11 では入力文中の「教師」を「生徒」に変換する.

表 2.11: 日本語側変換テーブルの適用例

入力文	私の姉は教師だ。
変換テーブル:C	教師
変換テーブル:A	生徒
一致する学習文対 (日本語側)	私の姉は生徒だ。

手順 2　学習文対に英語側の変換テーブルを適用

手順 1 と同じ変換テーブルの B と D を学習文対の英語側に適用し, 出力候補文を作成する. 表 2.12 では学習文対中の「student」を「teacher」に変換している.

表 2.12: 英語変換テーブルの適用例

一致した学習文対 (日本語側)	私の姉は生徒だ。
一致した学習文対 (英語側)	My sister is a student.
変換テーブル:B	student
変換テーブル:D	teacher
出力候補文	My sister is a teacher.

手順 3　最終的な出力文の決定

複数の出力候補文が得られた場合, 計算式 (2.31) に従って, 最終的な出力文を決定する. ここで P_m は言語モデルの確率である.

$$logP = logP_v + logP_m \tag{2.31}$$

図 2.3 に TDSMT の流れ図を示す.

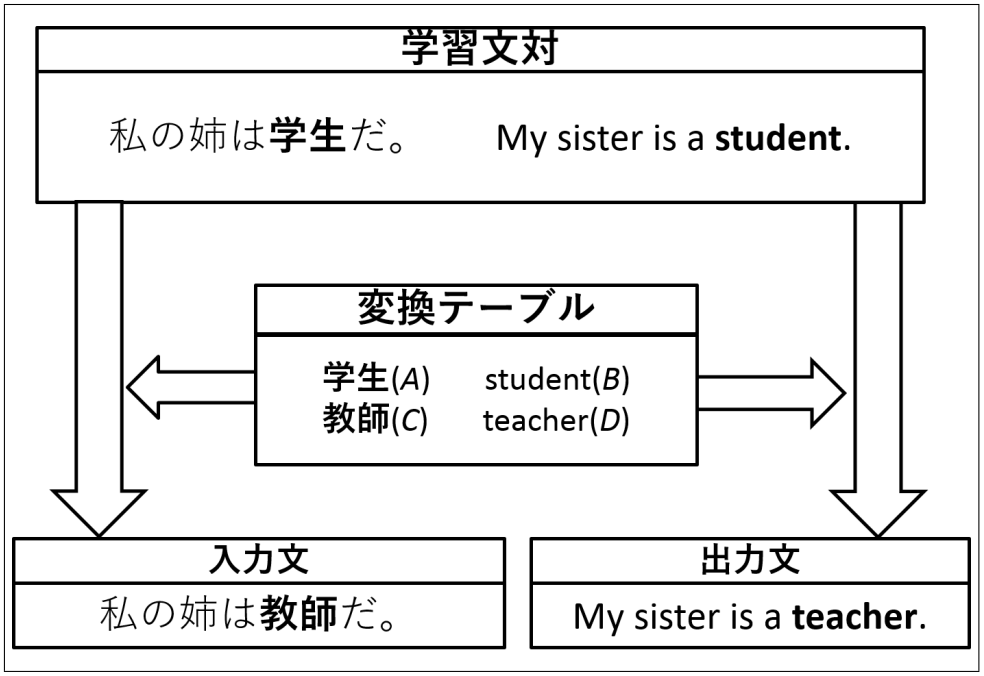

図 2.3: TDSMT の流れ図

2.3.4 変換テーブルの種類

変換テーブルには様々な種類が存在する. 具体的には以下のような変換テーブルが存在する.

- *ABCD* テーブル
- *ABAB* テーブル
- *ABCC* テーブル

・*ABCD* テーブル

ABCD テーブルは 2.3.2 節で述べた方法で作成する. *A* と *B* は対訳単語, *C* と *D* は対訳句である. *ABCD* テーブルの例を表 2.13 に示す.

表 2.13: ABCD テーブルの例

A	彼	*B*	he
C	学校 の 先生	*D*	school teacher

・*ABAB* テーブル

ABAB テーブルは対訳単語のみを利用する. *A* と *B*, そして, *C* と *D* は共に対訳単語である. *ABAB* テーブルの例を表 2.14 に示す.

表 2.14: ABAB テーブルの例

A	ヨーロッパ	*B*	Europe
C	映画	*D*	movie

・*ABCC* テーブル

ABCC テーブルは未知語の出力を目的とする変換テーブルである. *A* と *B* は対訳単語である. そして, *C* と *D* は同一の原言語の句である. *ABCC* テーブルの例を表 2.15 に示す.

表 2.15: ABCC テーブルの例

A	週	*B*	week
C	ピアノ の 勉強	*D*	ピアノ の 勉強

2.3.5 変換テーブルの問題点

TDSMT の問題点は, 誤った変換テーブルが存在する点である. 表 2.16 に誤った変換テーブルの作成過程を示す. 表 2.16 において, *X2* の変換テーブルの「C」は「髪」である. しかし,「D」 は 「dyed」である. ゆえに, TDSMT は *C*,*D* の訳が誤っているため, 置き換え不可能である変換テーブルが存在する.

そこで, 本研究では誤った変換テーブルを削除する方法を提案する.

表 2.16: 誤った変換テーブルの作成過程

<table>
<tr><td rowspan="2">学習文対 (1)</td><td colspan="4">日本語</td><td>私 は 英語 を 勉強し た 。</td></tr>
<tr><td colspan="4">英語</td><td>I studied English .</td></tr>
<tr><td rowspan="2">単語レベル
文パターン
(手順 2)</td><td colspan="4">日本語</td><td>X1 は X2 を X3 た 。</td></tr>
<tr><td colspan="4">英語</td><td>X1 X3 X2 .</td></tr>
<tr><td rowspan="2">学習文対 (2)</td><td colspan="4">日本語</td><td>彼女 は 髪 を 染め た 。</td></tr>
<tr><td colspan="4">英語</td><td>She had her hair dyed .</td></tr>
<tr><td rowspan="2">X3 の
変換テーブル
(手順 3)</td><td>A</td><td>英語</td><td>B</td><td colspan="2">English</td></tr>
<tr><td>C</td><td>髪</td><td>D</td><td colspan="2">dyed</td></tr>
</table>

第3章　変換テーブル選択

3.1　提案手法(変換テーブル選択)

本研究は誤った変換テーブルの削除を目的とする. 変換テーブルは *A* と *C*, また *B* と *D* の置き換えを想定している. 正しい変換テーブルは *A* と *C*, また *B* と *D* の文中での置き換えが可能である場合が多い. そこで本手法は, 文中における句の置き換え可否を検証し, 誤った変換テーブルを削除する.

本手法は, 分布仮説 [9] を利用し, 前後環境が等しい句は置き換えが文中において可能という考えに基づく. 前後環境は文中における句の前後の単語を利用する. 具体的には, 前後の単語の比較を学習文対内にて行う. そして, 一致する前後の単語の組み合わせが存在しなければ変換テーブルを削除する. 変換テーブル選択は日本語と英語を利用する. 以下は提案手法の手順である. なお, 変換テーブル選択は 2.3.4 節で述べた *ABCD* テーブルを対象とし, 提案手法を行なう.

手順 1 変換テーブルの句 (*A*,*B*,*C*,*D* に相当) の前後の単語を学習文対から取り出す

手順 2 手順 1 で取り出した前後の単語を比較する
　日本語選択 : *A* と *C* の前後の単語を比較する
　英語選択 : *B* と *D* の前後の単語を比較する

手順 3 合致する前後の単語が無ければ変換テーブルを削除する

なお, 提案手法はモノリンガルコーパスのみを利用して行うことが可能である.

3.1.1 日本語選択

表 3.1 の変換テーブルを表 3.2 の学習文対を利用して選択する場合を例に説明する.

表 3.1: 変換テーブルの例

A	嫌いだっ	*B*	disliked
C	追いかけ	*D*	made

表 3.2: 学習文対 (日本語側)

日本語文 (1)	文頭 私 は 彼 が **嫌いだっ** た 。
日本語文 (2)	文頭 私 は 彼女 を **追いかけ** た 。

表 3.1 の句 *A* , *C* の前後の単語を表 3.2 の学習文対内 (日本語文側) で比較する.

- 日本語文 (1) において “**嫌いだっ**” の前後の単語は「が・た」となっている
- 日本語文 (2) において “**追いかけ**” の前後の単語は「を・た」となっている

従って, 表 3.1 の変換テーブルを削除する.

3.1.2 英語選択

表 3.1 の変換テーブルを表 3.3 の学習文対を利用して選択する場合を例に説明する.

表 3.3: 学習文対 (英語側)

英語文 (1)	文頭 I **disliked** him .
英語文 (2)	文頭 I **made** after her .

表 3.1 の句 B , D の前後の単語を表 3.3 の学習文対内 (英語文側) で比較する.

- 英語文 (1) において “**disliked**” の前後の単語は「I・him」となっている
- 英語文 (2) において “**made**” の前後の単語は「I・after」となっている

従って, 表 3.1 の変換テーブルを削除する.

3.1.3 日本語&英語選択

日本語, 英語の両者で選択を行う.

第4章　実験

4.1　実験目的と方法

本実験の目的は変換テーブル選択による誤った変換テーブルの削除である. 実験方法はTDSMTで作成した変換テーブルに, 3.1.1, 3.1.2, 3.1.3節の変換テーブル選択を利用する. 評価は選択前と選択後の変換テーブルを比較して行う. 具体的には, 残った変換テーブルの数と精度によって提案手法である変換テーブル選択を評価する. また, 精度は人手評価により評価する.

4.2 実験データ

実験の評価は選択前の変換テーブルと選択後の変換テーブルの総数と精度を比較して行う. 変換テーブルの精度評価はCとDの訳の正誤によって判断する.

表 4.1 に変換テーブル作成と選択に用いた学習文対の総数を示す.

表 4.1: 変換テーブル作成と選択に使用した学習文対の総数

学習文対　159,998 対

本実験は学習文対 15,9998 対を変換テーブル作成に使用する. また, 本実験で使用する学習文対は日本語文と英語文の対である. 使用する学習文対は電子辞書などの例文より抽出した単文データである. 学習文対の例を表 4.2 に示す.

表 4.2: 学習文対の例

学習文対例 (1)	
日本語文	ピアノ の 勉強 に ヨーロッパ に 行く 。
英語文	Go to Europe to study the piano .
学習文対例 (2)	
日本語文	公園 は 川 まで 広がっ て いる 。
英語文	The park reaches to the river .
学習文対例 (3)	
日本語文	きょう は 時折 小雪 の ちらつく 寒い 一 日 だっ た 。
英語文	It was a cold day today with occasional light snowfall .

4.3 実験結果

4.3.1 変換テーブル数

変換テーブル数調査の結果を表 4.3 に示す. 最も多く変換テーブルが残った手法は日本語選択であった. 最も残った変換テーブルが少なかった手法は日本語 & 英語手法であった.

表 4.3: 変換テーブル数調査の結果

	変換テーブル数
選択前	4,308,398
日本語選択	2,851,461
英語選択	1,285,035
日本語 & 英語選択	986,031

4.3.2 変換テーブル精度

選択前と選択後の変換テーブルからランダムに 100 個抜き出し, 人手評価により精度を評価した. 精度は変換テーブルの C と D の訳の正誤によって判断した. 評価基準を以下に示す. 精度評価の結果を表 4.4 に示す.

○ : C, D の訳が正しい

△ : 不足している部分がある, または不必要な部分がある

× : C, D の訳が間違っている

表 4.4: 変換テーブル精度評価

	○	△	×
選択前	67	30	3
日本語選択	77	20	3
英語選択	85	13	2
日本語 & 英語選択	86	13	1

選択後の○の数は選択前と比べて増加した. また, 選択後の△の数は選択前と比べて減少した. よって, 提案手法を行なうことで変換テーブルの精度が向上することが確認できた. 以下に, それぞれの選択において評価した変換テーブルの具体例を示す.

変換テーブル具体例 : 日本語選択

○と評価した変換テーブルの例を表 4.5, 4.6, 4.7 に示す. △と評価した変換テーブルの例を表 4.8, 4.9, 4.10 に示す. × と評価した変換テーブルの例を表 4.11, 4.12, 4.13 に示す.

表 4.5: ○とした変換テーブル 1 (日本語選択)

変換テーブル			
A	見捨て	*B*	disowned
C	否認 し	*D*	deserted
日本語文 (1)	西川 君 の ドイツ 語 も **見捨て** た もの で は ない 。		
日本語文 (2)	その 交渉 へ の いかなる 関与 も **否認 し** た 。		

表 4.6: ○とした変換テーブル 2 (日本語選択)

変換テーブル			
A	水	*B*	water
C	東	*D*	east
日本語文 (1)	長野 県 から **東** を 関東 地方 と 呼ぶ 。		
日本語文 (2)	彼女 は 川 から **水** を 汲み上げ た 。		

表 4.7: ○とした変換テーブル 3 (日本語選択)

変換テーブル			
A	彼ら	*B*	they
C	フランケンシュタイン	*D*	Frankenstein's monster
日本語文 (1)	**彼ら** は 列 に 並ん だ 。		
日本語文 (2)	**フランケンシュタイン** は すごい 顔 を し て いる 。		

表 4.8: △とした変換テーブル 1 (日本語選択)

変換テーブル			
A	患者	*B*	patient
C	彼女 の 株	*D*	value of her stock
日本語文 (1)	**患者** の 容体 は 急 に 悪化 し た 。		
日本語文 (2)	**彼女 の 株** の 価値 が かなり 損なわ れ た 。		

表 4.9: △とした変換テーブル 2 (日本語選択)

変換テーブル			
A	サイレン	*B*	siren
C	彼女 に は 苛酷 な 運命	*D*	cruel destiny
日本語文 (1)	**サイレン** が 鳴っ た 。		
日本語文 (2)	**彼女 に は 苛酷 な 運命** が 待っ て い た 。		

表 4.10: △とした変換テーブル 3 (日本語選択)

変換テーブル			
A	壁	*B*	wall
C	関東 地方 一帯	*D*	whole Kanto area last night
日本語文 (1)	**壁** で 隣 の 家 と 続い て い ます 。		
日本語文 (2)	**関東 地方 一帯** で 朝 から 雪 が 降り 始め た 。		

表 4.11: × とした変換テーブル 1 (日本語選択)

変換テーブル			
A	やっ	*B*	did
C	得 なかっ	*D*	derived
日本語文 (1)	彼 は 人殺し も **やっ** た 。		
日本語文 (2)	宗教 から は 何 の 慰め も **得 なかっ** た 。		

表 4.12: × とした変換テーブル 2 (日本語選択)

変換テーブル			
A	割れ	*B*	broke
C	化膿 し	*D*	has come to a head
日本語文 (1)	氷 が **割れ** た 。		
日本語文 (2)	おでき が **化膿 し** た 。		

表 4.13: × とした変換テーブル 3 (日本語選択)

変換テーブル			
A	皿	*B*	plate
C	孤児 たち は 戦火	*D*	flames of
日本語文 (1)	**皿** の 縁 が 欠け た 。		
日本語文 (2)	**孤児 たち は 戦火** の 中 を 生き抜い た 。		

変換テーブル具体例 : 英語選択

○と評価した変換テーブルの例を表 4.14, 4.15, 4.16 に示す. △と評価した変換テーブルの例を表 4.17, 4.18, 4.19 に示す. × と評価した変換テーブルの例を表 4.20, 4.21 に示す.

表 4.14: ○とした変換テーブル 1 (英語選択)

変換テーブル			
A	変え	*B*	changed
C	上 の 方 へ 引き上げ	*D*	pulled up the top part of
英語文 (1)	I **changed** my plans .		
英語文 (2)	I **pulled up the top part of** my trousers .		

表 4.15: ○とした変換テーブル 2 (英語選択)

変換テーブル			
A	コンサート	*B*	concert
C	避暑 地	*D*	summer resort
英語文 (1)	The **concert** was broadcast on TV .		
英語文 (2)	The **summer resort** was almost deserted .		

表 4.16: ○とした変換テーブル 3 (英語選択)

変換テーブル			
A	煮え	*B*	cooked
C	高い ので 泳げ なかっ	*D*	too high to swim
英語文 (1)	The potatoes were **cooked** .		
英語文 (2)	The waves were **too high to swim** .		

表 4.17: △とした変換テーブル 1 (英語選択)

変換テーブル			
A	間違え	*B*	mistook
C	思いきっ て 言葉 を かけ	*D*	hazarded some remarks to break
英語文 (1)	I **mistook** the hospital for a hotel .		
英語文 (2)	I **hazarded some remarks to break** the monotony of the journey .		

表 4.18: △とした変換テーブル 2 (英語選択)

変換テーブル			
A	詩人	*B*	poet
C	私 の 息子 に	*D*	wife for my son
英語文 (1)	He is both a scholar and a **poet** .		
英語文 (2)	I'm searching for a **wife for my son** .		

表 4.19: △とした変換テーブル 3 (英語選択)

変換テーブル			
A	押す	*B*	Press
C	圧倒的 支持 を 得 た	*D*	The governor was supported overwhelmingly by
英語文 (1)	**Press** the button of the elevator .		
英語文 (2)	**The governor was supported overwhelmingly by** the citizens of the prefecture .		

表 4.20: × とした変換テーブル 1 (英語選択)

変換テーブル			
A	取り下げ	*B*	withdrawn
C	障害 物 が 何一つ	*D*	fair for our advance
英語文 (1)	The request was **withdrawn** .		
英語文 (2)	The way was **fair for our advance** .		

表 4.21: × とした変換テーブル 2 (英語選択)

変換テーブル			
A	妻	*B*	wife
C	ヘス	*D*	deputy
英語文 (1)	Jerry left his **wife** .		
英語文 (2)	Hitler made Hess his **deputy** .		

変換テーブル具体例 : 日本語 & 英語選択

○と評価した変換テーブルの例を表 4.22,4.23,4.24 に示す. △と評価した変換テーブルの例を表 4.25,4.26,4.27 に示す. × と評価した変換テーブルの例を表 4.28 に示す.

表 4.22: ○とした変換テーブル 1 (日本語 & 英語選択)

変換テーブル			
A	本	*B*	book
C	快い 驚き の ショック	*D*	pleasant shock of surprise
日本文 (1)	**本** を 買っ た 。		
日本語文 (2)	**快い 驚き の ショック** を 受け た 。		
英語文 (1)	I bought a **book** .		
英語文 (2)	I felt a **pleasant shock of surprise** .		

表 4.23: ○とした変換テーブル 2 (日本語 & 英語選択)

変換テーブル			
A	走る	*B*	run
C	小 づか い 銭 が 無くなっ た	*D*	have run short of pocket money
日本文 (1)	私 は **走る** 。		
日本語文 (2)	私 は **小 づか い 銭 が 無くなっ た** 。		
英語文 (1)	I **run** .		
英語文 (2)	I **have run short of pocket money** .		

表 4.24: ○とした変換テーブル 3 (日本語 & 英語選択)

変換テーブル			
A	必要	*B*	necessary
C	豪華 で 快適	*D*	plush and comfortable
日本文 (1)	食糧 増産 の 意味 から も 工業 の 発展 は **必要** で ある 。		
日本語文 (2)	この 車 の インテリア は **豪華 で 快適** で ある 。		
英語文 (1)	Extra care is **necessary** .		
英語文 (2)	The car's interior is **plush and comfortable** .		

表 4.25: △とした変換テーブル 1 (日本語 & 英語選択)

変換テーブル			
A	ウイスキー	*B*	whiskey
C	彼 は 次 の 歌手	*D*	next singer
日本語文 (1)	**ウイスキー** を 1 杯 もらおう 。		
日本語文 (2)	**彼 は 次 の 歌手** を 紹介 し た 。		
英語文 (1)	He drank a little of the **whiskey** .		
英語文 (2)	He announced the **next singer** .		

表 4.26: △とした変換テーブル 2 (日本語 & 英語選択)

変換テーブル			
A	問題	*B*	problem
C	合理 化 が 必要	*D*	rationalization
日本語文 (1)	時間 の **問題** で ある 。		
日本語文 (2)	我が 社 は 経営 の **合理 化 が 必要** で ある 。		
英語文 (1)	He referred to the **problem** of industrial pollution .		
英語文 (2)	Our company needs the **rationalization** of management .		

表 4.27: △とした変換テーブル 3 (日本語 & 英語選択)

変換テーブル			
A	港	*B*	harbor
C	秋田 犬 の 系統	*D*	Akita pedigree
日本語文 (1)	この 町 は **港** に 接し て いる 。		
日本語文 (2)	この 犬 は **秋田 犬 の 系統** に 属し て いる 。		
英語文 (1)	A ship slowly approached the **harbor** .		
英語文 (2)	The dog belongs to the **Akita pedigree** .		

表 4.28: × とした変換テーブル (日本語 & 英語選択)

変換テーブル			
A	任務	*B*	duty
C	道順	*D*	town hall
日本語文 (1)	故意 に 自分 の **任務** を 回避 し た 。		
日本語文 (2)	市役所 へ の **道順** を 尋ね た 。		
英語文 (1)	They are eager to be released from the **duty** .		
英語文 (2)	He asked directions to the **town hall** .		

具体例より, 日本語選択では合致した前後の単語に助詞が多く含まれていた. また, 英語選択では合致した前後の単語に冠詞が多く含まれていた. 対して, 日本語選択, 英語選択に共通して動詞が合致した前後の単語に含まれる場合は少なかった.

4.4 選択前に×とした変換テーブルの選択結果

表 4.4 の変換テーブルの精度調査結果より, × と評価した選択前の変換テーブルは 3 個存在した. × と評価した 3 個の変換テーブルを表 4.29, 4.30, 4.31 に示す.

表 4.29: × とした変換テーブル (1) (選択前)

変換テーブル			
A	馬	*B*	horse
C	すり に ご	*D*	against pickpockets

表 4.30: × とした変換テーブル (2) (選択前)

変換テーブル			
A	すっかり	*B*	completely
C	猛烈 な 駆け足 の あと 湯気 を	*D*	after a hard

表 4.31: × とした変換テーブル (3) (選択前)

変換テーブル			
A	建物	*B*	building
C	の 動議	*D*	to adjourn

選択前の 4 個の × とした変換テーブルの選択結果を調査した. 選択結果を表 4.32 に示す.

表 4.32: 選択前に × とした 3 つの変換テーブルの選択結果

	削除された数	残った数
日本語選択	3	0
英語選択	3	0
日本語 & 英語選択	3	0

表 4.32 より, 選択前の 3 個の × とした変換テーブルは日本語選択, 英語選択, 日本語 & 英語選択のいずれの手法においても全て削除された.

第5章　考察

5.1　変換テーブル数

表 4.3 より, 変換テーブル数は日本語選択の手法が英語選択の手法よりも多かった. 選択に使用する言語に依って変換テーブル数が異なる原因を, 表 5.1 の変換テーブルを選択する場合を例として考察する.

表 5.1: 変換テーブル例

A	サッカー	*B*	soccer
C	野球	*D*	baseball

5.1.1 英語選択

英語においては, 動詞が三単元や過去形のような活用の変化によりスペルが変化する. また, 名詞は単数形と複数形をとりスペルが変化する. 表 5.1 の変換テーブルを表 5.2 の英語コーパス文を利用して英語選択を行なう例を考える.

表 5.2: 英語コーパス文の例

英語文 (1)	I play **soccer** .
英語文 (2)	He plays **baseball** .

B である「soccer」の前後の単語は表 5.2 の英語文 (1) より「play」・「.」である. *D* である「baseball」の前後の単語は英語文 (2) より「plays」・「.」である. よって, 英語選択では表 5.1 の変換テーブルは *B* と *D* の前後の単語が合致しないことにより削除される. 従って, 英語選択においては, 単語の活用によりスペルが変化することで, 前後の単語が合致しない場合がある.

5.1.2 日本語選択

日本語文では句の前後に「の」,「は」,「を」のような助詞が存在する. 日本語選択においても, 前後の単語が合致した際の前後の組み合わせは助詞の組み合わせが多くみられた.

表 5.1 の変換テーブルを表 5.3 のコーパスを利用して英語選択を行なう例を考える. *A* である「サッカー」の前後の単語は表 5.2 の日本語文 (1) より「は」・「を」である. *C* である「野球」の前後の単語は日本語文 (2) より「は」・「を」である. よって, 日本語選択では表 5.1 の変換テーブルは *A* と *C* の前後の単語が合致することにより削除されない.

表 5.3: 日本語コーパス文の例

日本語文 (1)	私 は **サッカー** を する 。
日本語文 (2)	彼 は **野球** を する 。

よって, 日本語文の句の前後に助詞が存在することが日本語選択において最も多く変換テーブル数が残った原因である.

5.1.3　モノリンガルコーパスの利用

本手法はモノリンガルコーパスを追加利用することが可能である. モノリンガルコーパスによって, 選択に利用する文数を増やし, 英語選択後の変換テーブル数を増加させる. そして, 選択に利用する文数が増えることにより, 活用の変化に対応することが可能になる. 具体的には, 表 5.4 の追加文によって, 表 5.1 の変換変換テーブルは英語選択後に削除されなくなる.

表 5.4: 英語モノリンガルコーパス文の追加例

追加文	He plays **soccer** .
英語文 (1)	I play soccer .
英語文 (2)	He plays **baseball** .

5.2　精度の改善

表 4.4 より, 変換テーブル選択後の変換テーブルは選択前と比べ○が増え, × が減少している. よって, 提案手法は変換テーブルの精度を改善する. しかし, さらなる精度向上が見込める. 本考察では提案手法のさらなる精度向上を目指す改善策について述べる.

5.2.1　日本語選択

表 4.3 より, 日本語選択後の変換テーブル数は英語選択後よりも多い. しかし, 表 4.4 より, △と評価した変換テーブルの数は英語手法よりも多い. 日本語選択で△と評価した変換テーブルの例を表 5.5, 5.6 に示す.

表 5.5: △とした変換テーブル (1) (日本語選択)

変換テーブル			
A	乗り遅れ	*B*	missed
C	ニス を 薄く 塗っ	*D*	gave the desk a thin coat
日本語文 (1)	彼 は 机 に ニス を 薄く 塗っ た 。		
日本語文 (2)	２ 分 違い で 電車 に 乗り遅れ た 。		

表 5.5 において合致した前後の単語は 「に」・「た」である. また, 表 5.6 において合致した前後の単語は「が」・「た」である. つまり, 表 5.5 と表 5.6 の合致した前後の単語は共通して助詞を含む,

表 5.6: △とした変換テーブル (2) (日本語選択)

変換テーブル			
A	集まっ	*B*	gathered
C	軒 を 連ね て い	*D*	lined with shops
日本語文 (1)	通り に は 店 が 軒 を 連ね て い た 。		
日本語文 (2)	有志 の 者 が 集まっ た 。		

そこで, 改善策として, 前後 2 単語を比較する選択の手法を提案する. 前後 2 単語を比較することで, より正確に前後環境の利用が可能になると考えられる. 例として, 表 5.5 の変換テーブルを前後 2 単語を比較して選択を行なう. *A* である「乗り遅れ」の前後の 2 単語は, 「電車 に」・「た 。」 である. *C* である「ニス を 薄く 塗っ」の前後の 2 単語は, 「机 に」・「た 。」 である. よって, 表 5.5 は前後 2 単語を比較することで選択において削除される. 表 5.6 も同様に削除される. したがって, 前後 2 単語を比較することで△の変換テーブルをより削減できる.

5.2.2 英語選択

英語選択後の変換テーブルの中で△と評価した変換テーブルの合致した前後の単語を調査する. 英語選択で△と評価した変換テーブルの例を表 5.7, 5.8 に示す.

表 5.7: △とした変換テーブル (1) (英語選択)

変換テーブル			
A	魚	*B*	fish
C	すばやく ズボン	*D*	pair of pants
英語文 (1)	He swims like a fish .		
英語文 (2)	I quickly pulled on a pair of pants .		

表 5.8: △とした変換テーブル (2) (英語選択)

変換テーブル			
A	銀行	*B*	bank
C	５ 月 ６ 日 の	*D*	morning of May 6
英語文 (1)	She put the money in the bank .		
英語文 (2)	The vessel arrived at the port the morning of May 6 .		

表 5.7 において合致した前後の単語は 「a」・「.」である. また, 表 5.8 において合致した前後の単語は「the」・「.」である. 表 5.7 と表 5.8 の合致した前後の単語は共通して冠詞を含む,

日本語選択の改善策と同様に, 英語選択でも前後 2 単語を比較する改善策が有用である. 例として, 表 5.7 の変換テーブルを前後 2 単語を比較して選択を行なう. *B* である「bank」の前の 2 単語は, 「in」・「the」 , 後の単語は, 「.」 である. 後の単語が「.」のみであり, 後の単語が 2 単語に満たない場合は, 「.」の 1 単語のみを後の環境として利用する. *D* である「moring of May 6」の前の 2 単語は, 「port」・「the」 , 後の単語は 「.」 である. よって, 表 5.7 は前 2 単語, 後ろ 1 単語を比較することで選択において削除される. 表 5.8 も同様に削除される.

5.3 精度の改善

5.4 翻訳実験

提案手法を行った後の変換テーブルを利用して翻訳実験を行った. テスト文は 100 文を使用した. 翻訳には TDSMT を利用した.

5.4.1 カバー率

テスト文 100 文に対して, 得られた出力文の数を表 5.9 に示す. 従来手法の変換テーブルを利用した手法が最もカバー率が高かった. また, 提案手法後の変換テーブルを利用した手法はカバー率において, いずれも従来手法には及ばなかった. 提案手法後の変換テーブルを利用した手法がカバー率において, 従来手法に及ばなかった原因は, 提案手法による変換テーブルの減少が原因である.

表 5.9: 得られた出力文の数

	出力文数
従来手法	76
日本語選択	67
英語選択	74
日本語 & 英語選択	74

5.4.2 精度

表 5.9 の全ての手法で出力を得られた文を比較し, 精度を調査した. 精度調査の結果を表 5.10 に示す. 精度調査の評価基準を以下に示す.

○ **:** 従来手法より優れている

△ **:** 従来手法と同じ, または, 従来手法との優劣が一概に判断できない

× **:** 従来手法より劣っている

表 5.10: 出力文精度調査

	○	△	×
日本語選択	0	57	10
英語選択	1	60	6
日本語 & 英語選択	1	60	6

日本語選択を行った後の変換テーブルを使用した手法において△と評価した出力文を表 5.11, 5.12, 5.13 に示す. また, × と評価した出力文を表 5.14, 5.15, 5.16 に示す.

表 5.11: △とした出力文 (1) (日本語選択後の変換テーブル利用)

テスト文	ピアノ の 勉強 に ヨーロッパ に 行く 。
参照文	Go to Europe to study the piano .
従来手法	We go to a Europe in a ピアノ の 勉強 .
日本語選択	We go to a Europe in a ピアノ の 勉強 .

表 5.12: △とした出力文 (2) (日本語選択後の変換テーブル利用)

テスト文	この 季節 として は よい 天気 だ 。
参照文	It's good weather for this time of year .
従来手法	The この 季節 として weather good .
日本語選択	The この 季節 として weather good .

表 5.13: △とした出力文 (3) (日本語選択後の変換テーブル利用)

テスト文	公園 は 川 まで 広がっ て いる 。
参照文	The park reaches to the river .
従来手法	The park is spread until the river .
日本語選択	The park is spread till river .

表 5.14: ×とした出力文 (1) (日本語選択後の変換テーブル利用)

テスト文	友達 と ハイキング に 行き まし た 。
参照文	I went hiking with some friends .
従来手法	I went for hiking with a friend .
日本語選択	He was 行き まし to 友達 と ハイキング .river .

表 5.15: ×とした出力文 (2) (日本語選択後の変換テーブル利用)

テスト文	私 は その 仕事 に 慣れ て い ない 。
参照文	I am not used to the task .
従来手法	I am the not used to work .
日本語選択	I is その 仕事 に 慣れ て い .

表 5.16: ×とした出力文 (3) (日本語選択後の変換テーブル利用)

テスト文	彼 は 彼女 に 慰め の ことば を かけ た 。
参照文	He spoke comforting words to her .
従来手法	He spoke a few remark of comfort to her .
日本語選択	He hung her a 慰め の ことば .

英語選択を行った後の変換テーブルを使用した手法において○と評価した出力文を表5.17に示す. △と評価した出力文を表5.18, 5.19, 5.20に示す. また, × と評価した出力文を表5.21, 5.22, 5.23に示す.

表5.17: ○とした出力文 (英語選択後の変換テーブル利用)

テスト文	この 分 を 3 個 ください 。
参照文	Please give me three of these .
従来手法	Please the minutes 3 個 .
英語選択	Please pieces the この 分 into three slices .

表5.18: △とした出力文 (1) (英語選択後の変換テーブル利用)

テスト文	油井 から 石油 が 吹き出し た 。
参照文	Oil gushed out of the well .
従来手法	My delivers broke out in oil .
英語選択	My delivers broke out in oil .

表5.19: △とした出力文 (2) (英語選択後の変換テーブル利用)

テスト文	彼 は その 白馬 を 追いかけ た 。
参照文	He went after the white horse .
従来手法	He chased その 白馬 を .
英語選択	He chased その 白馬 を .

表5.20: △とした出力文 (3) (英語選択後の変換テーブル利用)

テスト文	彼 の 性質 は ひねくれ て いる 。
参照文	He has an uneven disposition .
従来手法	His nature is distorted views .
英語選択	His nature is twisted .

表 5.21: ×とした出力文 (1) (英語選択後の変換テーブル利用)

テスト文	日本 は 太陽熱 温水 技術 で アメリカ に 水 を あけ た 。
参照文	Japan has opened up a lead over the US in solar water heater technology .
従来手法	Japan drilled a America in water to the 太陽熱 温水 技術 .
英語選択	Japan drilled a アメリカ に 水 in the 太陽熱 温水 技術 .

表 5.22: ×とした出力文 (2) (英語選択後の変換テーブル利用)

テスト文	彼 の 話 に は 深み が ある 。
参照文	There is something deepin what he said .
従来手法	What his said is a depth in story .
英語選択	There are depth contained in 彼 の 話 .

表 5.23: ×とした出力文 (3) (英語選択後の変換テーブル利用)

テスト文	兄 は 体力 で 弟 に 劣っ て いる 。
参照文	The elder brother is inferior to the younger brother in terms of physical strength .
従来手法	My brother is inferior to 体力 で 弟 .
英語選択	My brother is strength in 弟 に 劣っ .

英語 & 日本語選択を行った後の変換テーブルを使用した手法において○と評価した出力文を表 5.24 に示す. △と評価した出力文を表 5.25, 5.26, 5.27 に示す. また, × と評価した出力文を表 5.28, 5.29, 5.30 に示す.

表 5.24: ○とした出力文 (日本語 & 英語選択後の変換テーブル利用)

テスト文	この 分 を 3 個 ください 。
参照文	Please give me three of these .
従来手法	Please the minutes 3 個 .
日本語 & 英語選択	Please pieces the この 分 into three slices .

表 5.25: △とした出力文 (1) (日本語 & 英語選択後の変換テーブル利用)

テスト文	その 翌日 の バス の 切符 を 買っ た 。
参照文	He booked himself for the following day's bus .
従来手法	I bought a ticket from the 翌日 の バス .
日本語 & 英語選択選択	I bought a ticket from the 翌日 の バス .

表 5.26: △とした出力文 (2) (日本語 & 英語選択後の変換テーブル利用)

テスト文	北海道 の 開拓 民 は 荒れ地 を 開い て いっ た 。
参照文	The settlers opened up the wasteland in Hokkaido .
従来手法	The Hokkaido 開拓 民 開い て いっ the wilderness .
日本語 & 英語選択選択	Hokkaido の 開拓 民 開い て いっ the wilderness .

表 5.27: △とした出力文 (3) (日本語 & 英語選択後の変換テーブル利用)

テスト文	最後 に 皆 で 校歌 を うたっ た 。
参照文	At the end we all sang our school song together .
従来手法	She sang a 皆 で 校歌 in the last .
日本語 & 英語選択選択	She sang a 皆 で 校歌 in the last .

表 5.28: ×とした出力文 (1) (日本語 & 英語選択後の変換テーブル利用)

テスト文	中国人 が 印刷 技術 を 発明 し た 。
参照文	The Chinese invented printing .
従来手法	The Chinese invented movable in the printing technique .
日本語 & 英語選択選択	A Chinese invention the 印刷 技術 .

表 5.29: ×とした出力文 (2) (日本語 & 英語選択後の変換テーブル利用)

テスト文	川 で は 濁流 が 渦巻い て い た 。
参照文	The muddy stream was whirling around in the river .
従来手法	The muddy were whirling in a the river .
日本語 & 英語選択選択	The は 濁流 were whirling in the river .

表 5.30: ×とした出力文 (3) (日本語 & 英語選択後の変換テーブル利用)

テスト文	当地 で は 家賃 が インフレーション より 急速 に 上がっ て いる 。
参照文	Rents are rising faster than inflation here .
従来手法	Anti-Japanese am going up to インフレーション より 急速 for the rent .
日本語 & 英語選択選択	My 当地 で is 家賃 が インフレーション than 急速 に 上がっ .

出力文の精度が従来手法と比べ低下した原因は, 未知語出力用の *ABCC* テーブルの使用である. 提案手法が従来に使用していた *ABCD* テーブルを削除した場合, 提案手法後の変換テーブルを利用した翻訳では, 従来とは異なる学習文対を使用する例が存在した. 表 5.31 は従来手法の例である. 表 5.32 は提案手法後の変換テーブルを使用した手法の例である. なお, 表 5.32 の出力文は日本語選択後, 英語選択後, 日本語 & 英語選択後の変換テーブルを利用した手法すべてにおいて共通していた.

提案手法である変換テーブル選択は, 表 5.31 の *ABCD* テーブルを削除した. よって, 表 5.32 では表 5.31 とは異なる学習文対を使用している. また, 表 5.32 では従来手法が使用していない *ABCC* テーブルを使用している. つまり, *ABCD* テーブルの削除による, 異なる学習文対の使用が *ABCC* テーブルの使用を引き起こす. 結果, 提案手法後の変換テーブルを使用した場合, 出力文には未知語が増加し従来手法と比べ精度が低下する.

表 5.31: ABCD 使用例

従来手法				
テスト文	川 で は 濁流 が 渦巻い て い た 。			
学習文対	嵐 で 帆 が 裂け た 。			
	The sail split in the storm .			
変換テーブル (*ABCD*)	A	裂け	B	split
	C	渦巻い て い	D	were whirling
出力文	The muddy were whirling in a the river .			

表 5.32: ABCC 使用例

提案手法後の変換テーブル使用				
テスト文	川 で は 濁流 が 渦巻い て い た 。			
学習文対	空 に 星 が きらめい て い た			
	The stars scintillated in the sky .			
変換テーブル (*ABCC*)	A	星	B	stars
	C	は 濁流	D	は 濁流
出力文	文頭 The は 濁流 swirled in the river .			

5.5 変換テーブルの閾値

TDSMT は変換テーブルに確率を付与し, 順位付けを行う. そして, 順位が低い変換テーブルは閾値を用いた枝刈りにより削除する. また, 確率付与には IBM Model を利用している.

本実験では枝刈り後の変換テーブルに対して提案手法を行った. 枝狩り前の変換テーブルを用いて提案手法を行った場合, 選択後の変換テーブル数は枝刈り後を選択した場合よりも多くなる. しかし, 誤った変換テーブルが残る可能性が高い.

第6章　おわりに

機械翻訳において, 相対的意味論に基づく変換主導型統計機械翻訳 TDSMT が提案されている. TDSMT は変換テーブルを用いて, 学習文対を変換し, 出力文を作成する. 変換テーブルは学習文対 (パラレルコーパス) から自動作成する. しかし, 自動作成のため誤った変換テーブルが存在する. そこで, 本研究は誤った変換テーブルの削除を目的とした.

本研究では提案手法として, 前後環境を利用して誤った変換テーブルを削除する方法を提案した. 提案手法によって変換テーブルの精度の向上を確認した.

今後の課題として, 精度向上のために前後 2 単語を比較する変換テーブル選択, そして, 数の向上のためにモノリンガルコーパスの利用が挙げられる. また, 翻訳実験においては, 学習文対量の増加による *CDAB* テーブルの増加, また, 未知語出力用変換テーブルの不使用によって従来手法と提案手法の出力の違いが明確に表れる.

謝辞

本研究を進めるにあたり, 研究の説明や論文の書き方など様々なご指導を頂きました鳥取大学工学部電気情報系工学科自然言語処理研究室の村上仁一准教授に心から御礼申し上げます. また, 本研究を進めるにあたり, 御指導, 御助言を頂きました, 村田真樹教授に心から御礼申し上げます. また, 同じ班に所属されていた今仁先輩, 金子先輩, 中村先輩をはじめとする自然言語処理研究室の皆様へ心から感謝の気持ちと御礼を申し上げたく, 謝辞にかえさせて頂きます.

参考文献

[1] 安場裕人, 村上仁一 (2018). “変換主導型翻訳の提案”. 自然言語処理学会第 24 回年次大会.

[2] Philipp Koehn (2005). “Europarl: A Parallel Corpus for Statistical Machine Translation”. *MT Summit*, pp.79?86.

[3] Peter F.Brown, Stephen A.Della Pietra, Vincent J.Della Pietra, Robert L.Mercer (1993). “The mathematics of statistical machine translation: Parameter Estimation”. *Computational Linguistics*.

[4] 西尾聡一郎 (2016). “パターンに基づく統計翻訳における文パターン確率の考察”. 平成 27 年度 卒業論文, pp.3-16.

[5] 力久 剛士 (2015). “レーベンシュタイン距離を用いた翻訳精度の向上”. 平成 26 年度 卒業論文, pp.3-15

[6] Philipp Koehn, Marcello Federico, Brooke Cowan, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst (2007). “Moses: Open Source Toolkit for Statistical Machine Translation”. *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp.177-180.

[7] 中村 勇太 (2019). “ 相対的意味論に基づく変換主導型パターンベース統計機械翻訳（TDPBSMT）の提起”. 平成 30 年度 卒業論文.

[8] Franz Josef Och, Hermann Ney (1996). “A Systematic Comparison of Various Statistical Alignment Models”. *Computational Linguistics*, 29(1), pp.299-314.

[9] Zellig S.Harris (1954). “Distributional structure”. *Word*, Vol. 10, No.23, pp.146-162.